DOCUMENT RESUME

ED 202 239 FL 012 321

AUTHOR Gaies, Stephen J.
TITLE Language Testing and the Notion of "Authenticity."
PUB DATE 80
NOTE 13p.; Paper presented at the Kentucky Foreign Language Conference (April 24-26, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cloze Procedure; Context Clues; Interviews; *Language Tests; Objective Tests; Postsecondary Education; Secondary Education; *Second Language Instruction; *Test Validity; Writing (Composition)

ABSTRACT

As a reaction to the popularization of so-called "discrete point" tests of language proficiency, that is, tests that evaluate control of a sample of specific linguistic items, a number of testing formats which provide a more global, integrative measure of language proficiency have been proposed or revived in recent years. Some of these measures are: oral interviews, written compositions, and cloze tests. Among the advantages often cited for this latter approach is the greater face validity of such tests; they require subjects to perform tasks which more closely mirror language use in the "real world." The thesis of this paper is that claims about face validity, or "authenticity," to borrow a term Widdowson (1976, 1978) uses in connection with the nature of learning materials, are often based on an oversimplified notion of face validity and are thus to a large degree specious. An examination of a number of widely used "integrative" testing formats will reveal that face validation encompasses at least four variables and that a testing format which would be fully face valid would be impractical and uneconomical for most settings. The conclusion of the above argument is that the only truly authentic language tests are those which take place outside the context of a language instruction setting. The implication of this for statistical approaches to test validation are outlined. (Author/AMH)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

# Language testing and the notion of "authenticity"[1]

Stephen J. Gaies
University of Northern Iowa

The field of language teaching has often found it useful to discuss issues and trends in terms of dichotomies. We talk of methods as either inductive or deductive; of syllabuses as either structurally- or situationally-ordered; of learners as either instrumentally or integratively motivated. Generally speaking, though, we recognize that terms such as these represent the ends of a continuum, and that methods, syllabuses and learners fall somewhere within that continuum.

Language testing has used its share of such terms. One of the most well-known distinctions is the one between "discrete point" and "integrative" testing. One way to appreciate the difference between the two is to examine the role which "context" plays in the two approaches. In discrete point testing, the focus at any one time is on learners' ability to demonstrate knowledge of a single item or element of the target language. The ideal test item is the one which shows unambiguously whether or not the learner has mastered the element in question. For this reason, it is claimed, elements of the target language are best tested "out of context"; in this way learners cannot arrive at the correct answer except by actually "knowing" the specific point being tested. Examples 1 and 2 illustrate two test items representing different levels of "de-contextualization". Whereas in the first version of the item there are a number of possibilities for arriving at the correct response, in the second no such possibilities exist.

ED202239

FL012321

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Stephen J. Gaies

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Example 1.

Directions: Read the following short conversation. Then circle the letter of the phrase which means the same thing as the underlined word.

Speaker 1: John, you look awfully tired. Do you feel all right?
Speaker 2: I guess so. I hardly got any sleep last night. I've had trouble keeping my eyes open all day.
Speaker 3: You ought to take a nap when you get home.

a. kind of medicine
b. brief sleep
c. leisurely walk
d. relaxing bath

Example 2.

Directions: Circle the letter of the word which is closest in meaning to the underlined word or phrase.

nap

a. hold
b. sleep
c. walk
d. eat

Integrative tests, on the other hand, are those in which part of what is being assessed is the ability of learners to use context as a tool in the performance of given language tasks. Examination of "cloze" tests, dictations, and direct measures of language proficiency such as oral interviews and written compositions will show that learners have a variety of clues available to them as they attempt to construct, reconstruct, or interpret oral or written language. Because they reflect very clearly the importance of context in language activity, cloze tests are a good example of the integrative approach to language testing. The contextualized nature of integrative testing formats is also at the heart of the claim that such formats require learners to perform tasks which, in comparison to discrete point items, more closely mirror ordinary language activity. It is certainly true that if we consider language as it is used in the "real world," we can find without any difficulty language activities analogous to dictation, oral interviews, and various tasks in written production. Thus, it is widely argued that integrative tests have "face validity" and are, at least in that respect, superior to discrete point testing.

Our concern here is not with the fact that discrete point tests lack face validity; the discrete point approach has been justified on other grounds, and even its most ardent advocates concede that the task involved in an item such as Example 3 does not reflect in any demonstrable way what speakers do in ordinary language use. Rather, it is the claim that integrative tests <u>have</u> face validity that we wish to examine in these pages. The

---

Example #3

Directions: Write the number of the sound that is the same as the sound represented by the letters omitted.

ITEM X: "Light the fire with dry w--d."

1. F<u>oo</u>d.
2. C<u>u</u>t.
3. L<u>oa</u>d.
4. W<u>a</u>lk.
5. Understood.[2]

---

theme of these pages is as follows: that claims made about face validity are often based on an oversimplified notion of face validity and are thus to a large degree specious.

In making this argument, we should first define "validity" as a general term and "face validity" as a specific concept. Validity has to do with whether a testing instrument in fact measures what it claims to measure. In the simplest possible terms, we can say that a test of proficiency in a language ought to involve examinees in language-related tasks. Tests which claim to measure language ability, but which involve, let us say, an inordinate degree of mathematical reasoning ability or some other non-linguistic skill (or which demand previous knowledge of some specific subject matter), lack validity.

It is easy to recognize testing instruments which are grossly invalid. It is relatively more difficult to measure validity, since tradition has offered a number of alternatives. There are,

first of all, statistical approaches to test validation, in which test-makers indicate a relationship between performance on their test and performance on some other, already-validated instrument (concurrent validity) or performance in some real-world task (predictive validity). This is perhaps the most widely-used and well-established approach to language test validation. Another approach is to show that the test requires subjects to demonstrate precisely the skills and knowledge that linguistic theory has identified as central to knowing a language. Validation of this sort, which is referred to as "construct validation," has been identified as the most difficult, but ultimately the most fruitful, approach to the issue of language test validity.

In contrast to the approaches to validity just discussed, "face validity" has to do with the nature of the testing instrument as it is perceived by the people to whom it is administered. Tests that strike subjects as unrelated to what they are supposed to be measuring, as well as those which for one reason or another annoy, insult, or embarrass subjects, or which place inordinate pressure on them, may be said to lack face validity. It is sometimes impossible to anticipate the way in which subjects will react to a particular testing format. As a general rule, however, the more closely the required tasks reflect what individuals ordinarily do in language activity, the more likely the test will be assumed to be face valid.

By and large, this is as much as is usually said on the subject of face validity: whereas tests have some degree of statistical validity, they are usually classified as having or not having face validity. In fact, however, face validity is not an

"either/or" matter. Rather, language tests mirror "real world" language activity to varying degrees. We may understand this point better if we relate the concept of face validity to the notion of "authenticity".

"Authenticity" is a term originally used by Henry Widdowson[3] in connection with language teaching materials. Authenticity is not an inherent quality: that is, language learning material is not in itself either authentic or inauthentic. It is authentic only under certain circumstances: namely, when we cause learners to respond to it in the way users of the language would ordinarily do so. Thus, if we have our learners listen to a tape of an actual news broadcast and then quiz them on selected vocabulary items from the tape, we are asking them to respond to that language material inauthentically. This is not necessarily a bad thing--indeed, at beginning levels of instruction it may be the most useful if not the only course to pursue. Widdowson's point is that the use of "genuine" material--that is, actual samples of language use--rather than material contrived or doctored for pedagogical purposes is not especially important in itself. He goes on to state, however, that at later stages, the goal of developing communicative ability in learners can be reached only by providing learners with the opportunity to respond authentically to language: to recognize the conventions of language use and to respond appropriately.

If we transfer the notions of authenticity to the field of language testing, we will perhaps be better able to consider claims made for the face validity of language proficiency tests. We will find, in fact, that for a test to involve learners in activities analogous to those we perform in ordinary language

use, the tasks must satisfy no less than four criteria:

1. The task. The tasks themselves must of course have some analog in normal language activity. Although it is not always easy to recognize violations of this principle, two recurrent problems can be cited:

a. tasks which violate the conventions of language use specifying narrow and in some cases unnatural modes of response: for example, the requiring of full-sentence answers to questions rather than giving credit for more natural elliptical responses;[4] and

b. tasks which place inordinate demands on language-related abilities such as memory. Such a task is presented in Example 4.

---

Example #4

Directions: You will hear four words repeated one after another. Listen carefully to determine which words are the same. Write the number of the words that are the same.

(EXAMINER reads: 1. "share"
2. "share"
3. "chair"
4. "share")

On the answer sheet the STUDENT writes the numbers 1, 2, and 4.[5]

---

Related to this is the problem of tests which require subjects to perform tasks which do not involve language ability exclusively or primarily. Tests of general aptitude or specific abilities such as mathematical reasoning or map reading not only are invalid on other grounds; to the extent that these are perceived by subjects themselves as invalid, they also lack face validity.

2. The testing context. It is widely agreed that language testing must involve a willing suspension of belief on the part of examinees. As Clark points out, "it is probably futile to hope that the affective components of a formal testing situation will ever closely approach those of the real-life situations which the test attempts to reflect."[6] Perren's remarks on this point, which Clark cites, are worth repeating here:

> ...both participants know perfectly well that it is a test and not a tea party, and both are subject to psychological tensions, and what is more important, to linguistic constraints of style and register thought appropriate to the occasion by both participants.[7]

The unnaturalness of the test situation can be minimized but never eliminated. What we hope to do is to identify those things that aggravate the unrealistic nature of the testing context. Examples of these are:

a. tests of language dominance which require an examinee to answer the same question in two languages, using on immediately after the other, or in which language switching is done on the basis of the examiner's directions alone; and

b. tests which ask the examinee to supply information which is obvious on the basis of extralinguistic knowledge (for example, "Are you sitting or standing?").

3. The evaluation of performance. We often devise tests which are fair approximations of ordinary language use, but for which our criteria for assessment violate the conventions by which language proficiency is normally judged. Examples of such criteria are: degree of verbal initiative assumed by an interviewee, use

of full-sentence answers, quantity of verbal output, and the syntactic complexity of verbal output. We should hasten to add, of course, that there may be many instances in which we would want to know about our learners' performance on these criteria, since such information may enable us to chart progress in second language development. It is the use of any of these criteria by itself as an index of proficiency in the target language that will cause our instrument to lack face validity.

4. The "stakes". The final criterion by which the authenticity of a test must be judged is the issue of what is at stake. To what degree does a subject's performance on a task relate to the normal system of rewards and penalties which operates in ordinary language activity? Think, for example, of the oral interview in its relationship to ordinary semi-formal verbal interaction. In the case of the oral interview as a language test, one participant (the examiner) seeks to establish the ability of the other (the examinee) to communicate information intelligibly and with reasonable linguistic accuracy. Essentially, though, what the examinee says need be neither particularly truthful nor particularly interesting to the examiner. In the real-world analog, however, those are precisely the primary standards by which the quality of verbal interaction is judged. We do not <u>normally</u> engage in verbal interaction to listen to our co-participants tell us what we already know or what we are not interested in hearing.

The issue of the stakes involved for a participant in language testing is especially troublesome in diagnostic testing, where, it might be argued, the examinee has little to gain or lose. But indeed, as a general rule, we might state that any testing situation

in which the examinee is relatively unaware of what the stakes are--or where the stakes are extremely low--violates a fundamental convention of language use and makes authenticity in the relationship of the examinee to the language activity impossible.

If we consider test validity in terms of these criteria, we reach the conclusion that no test can be fully face valid. Rather, as was already stated, any test will be only relatively valid or invalid. This should not be cause for dismay, however, since no test is invulnerable to criticism; there are other kinds of validity, as well as demands for reliability and practicality, all of which no single instrument can fully meet. Indeed, we have long accepted the fact that language testing, to whatever degree it involves science, also involves the art of compromise.

In terms of face validity, however, we must reconcile ourselves to the view that the only fully valid language tests are those which take place outside the context of instructional settings and which we label not language tests, but language use. We need to recognize this and recognize also its implications for further attempts to develop useful language tests. One such implication is that our efforts to validate tests ought to be geared as much as possible toward establishing the relationship between performance on individual tests with performance by the same subjects in real-world language tasks. To attempt to validate a test by saying that performance on it correlates with performance on some other instrument is, in view of the argument presented above, something of a circular statement. Quite simply, such statements tell us nothing about what we most need to know: namely, to what degree the convergence of

learner, task, context, and the basis of assessment used in the test we have developed allows us to observe examinees involved in authentic language activity. This, I would contend, remains the enormous and enormously important challenge facing language test developers today.

---

Notes

[1]This is a revised version of a paper given at the 33rd Annual University of Kentucky Foreign Language Conference, held in Lexington, Kentucky, April 24-26, 1980.

[2]The example is from Robert Lado, Language Testing: The Construction and Use of Foreign Language Tests (London: Longmans, 1975), p. 102.

[3]Henry G. Widdowson, "The Authenticity of Language Data, in On TESOL '76, ed. John H. Fanselow and Ruth H. Crymes (Washington, D.C.: Teachers of English to Speakers of Other Languages, 1976), pp. 261-70. The concept of "authenticity" is also discussed in Henry G. Widdowson, Teaching Language as Communication (Oxford University Press, 1978), pp. 79-81.

[4]This problem and a number of others are also discussed in Thomas G. Dieterich, Cecilia Freeman, and Jo Ann Crandall, "A Linguistic Analysis of Some English Proficiency Tests," TESOL Quarterly, 13 (1979), 535-50.

[5]This example is based on one in Lado, p. 55.

[6]John L.D. Clark, "Psychometric Considerations in Language Testing," in Approaches to Language Testing, ed. Bernard Spolsky (Arlington, Va.: Center for Applied Linguistics, 1978), p. 24.

[7]G.E. Perren, "Testing Ability in English as a Second Language: 3. Spoken Language," English Language Teaching, 22 (1967), 22-9.